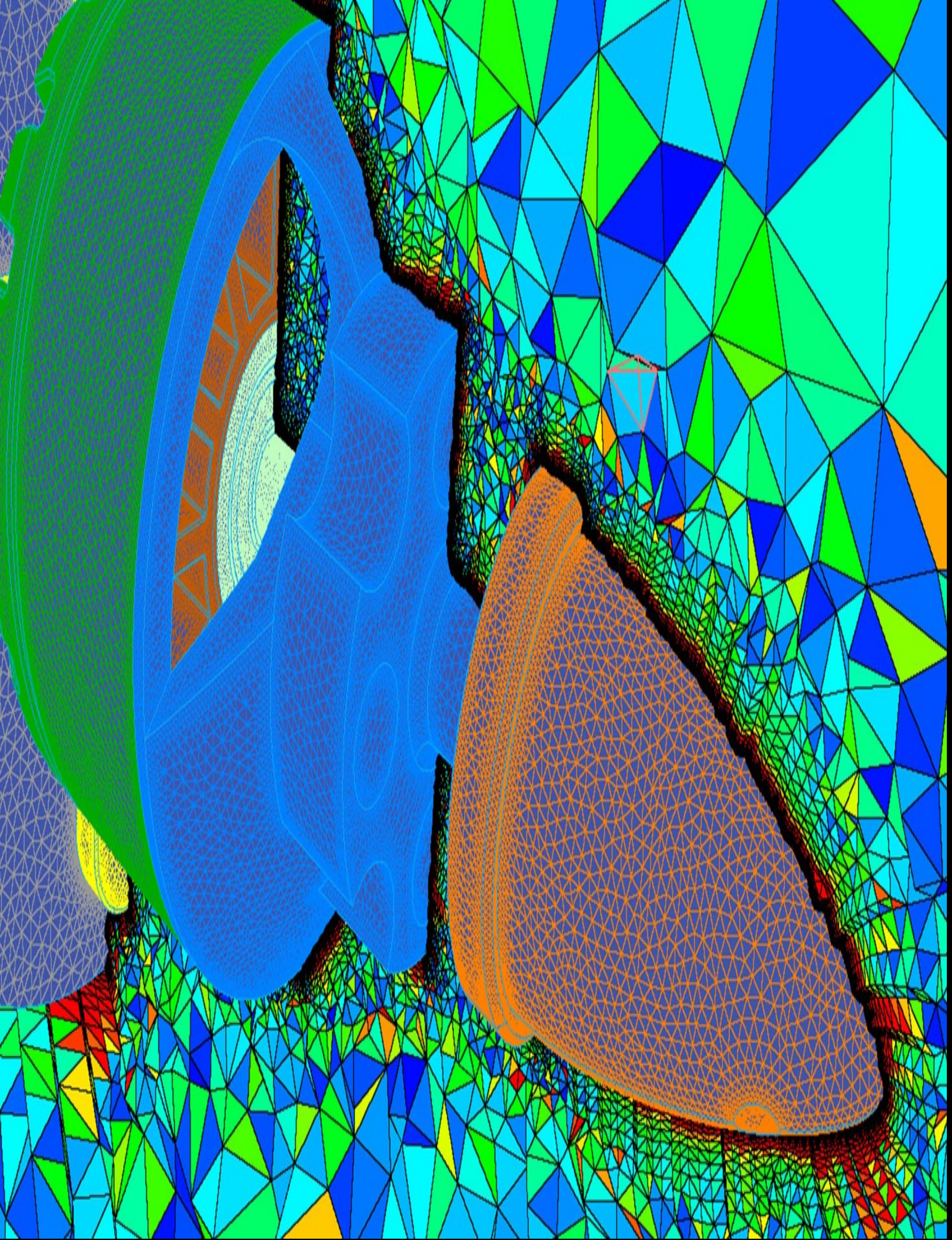


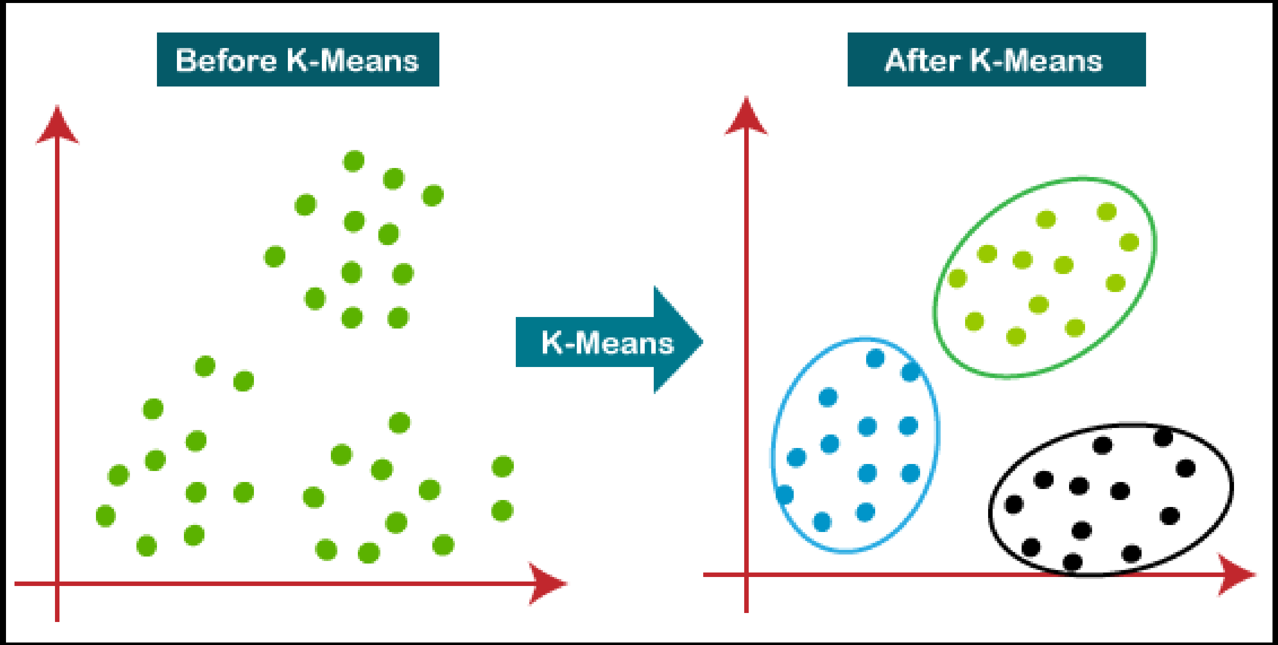
MÉTODO NUMÉRICOS

Aula 1 –Introdução ao Erro

Curso de Ciência da Computação
Dr. Rodrigo Xavier de Almeida Leão
Cientista de Dados

QS (toneladas)	L (milhões de reais)
0	0,00
100	0,04
200	0,16
400	0,64
500	1,00
600	1,43
800	2,55

Com essas informações e com o auxílio de um software capaz de gerar a relação matemática dessas duas variáveis, observou-se que a curva que descreve o comportamento é $L = \ln(QS+1)\ \ 10^6 + 4QS^2/10^6 - QS/10^5$. Então, o seu gerente solicita algumas informações: primeiramente (na Seção 1.1), ele necessita que você determine a quantidade de silicone que deve ser produzida para alcançar um lucro de R$ 600.000,00. Depois (na Seção 1.2), ele deseja a sistematização conceitual dos procedimentos necessários para outras formas de determinação da quantidade de silicone produzida num valor fixado de lucro, considerando o modelo apresentado anteriormente, e se é possível determinar a quantidade de silicone que maximiza o lucro da empresa. Por fim (na Seção 1.3), ele requisita que sejam desenvolvidos algoritmos para a implementação computacional da função "quantidade de silicone e lucro", considerando os conceitos apresentados a você na segunda etapa (na Seção 1.2).



Prof. Dr. Rodrigo Xavier de Almeida Leão

BINÁRIO <-> DECIMAL

$$N = \pm a_n 2^n + a_{n-1} 2^{n-1} + \ldots + a_1 2^1 + a_0 2^0$$

$$\text{Com: } \begin{cases} a_n = 1 \\ a_i = 0 \quad \text{ou} \quad 1, \quad 0 \le i < n \end{cases}$$

Assim, se quiséssemos converter $(100000)_2$, número de base 2, para base 10, analisaremos que temos 6 dígitos, $n = 5$, e identificamos os termos a_i por: $a_5 = 1$, $a_4 = 0$, $a_3 = 0$, $a_2 = 0$, $a_1 = 0$, $a_0 = 0$. Em seguida, utilizamos a fórmula da expansão binária:

$$(100000)_2 = 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 = 32$$

Se quisermos realizar a operação oposta, ou seja, conversão de um sistema decimal para binário, necessitaremos do método das divisões sucessivas sobre a base 2 até que o último quociente seja menor que a base, no caso, 1.

```
20|2
0   10|2
     0   5|2
         1   2|2
             0   1
```

Assim, tem-se que $(20)_{10} = (10100)_2$.

IMPLEMENTAR FUNÇÃO DE CONVERSÃO DECIMAL <-> BINÁRIO

75 PTS

Prof. Dr. Rodrigo Xavier de Almeida Leão

Formato ponto flutuante

Ponto flutuante é a representação dos números reais empregada em máquinas. Basicamente, esse número é composto por três partes: sinal, mantissa e expoente, e é identificado por:

$$m = \pm \quad ,d_1 d_2 d_3 \ldots d_t \cdot \beta^e$$

Sendo: d_i: dígitos da parte fracionária, $d_1 \neq 0$; β : base (2 - binário; 10 - decimal); t : número de dígitos na mantissa; e : expoente inteiro. De forma simplificada: $F(\beta, t, e_{\min}, e_{\max})$.

Podemos representar o número 43,6, em base decimal, com 4 dígitos na mantissa por: $x = 43,6$; $\beta = 10$; $t = 4$, cuja representação será $x = 0\ 4360 \cdot 10^2$.

Erros na representação dos números

Os números reais formam um conjunto infinito de números, então imagine a quantidade de números que podemos ter num intervalo entre 0 e 1. Por outro lado, a representação deles no sistema de ponto flutuante é finita, pois a faixa dos expoentes é limitada, ou seja, $e_{min} < e < e_{max}$. Observe: dado que $F(10,2,-5,5)$ e que se deseja fazer a operação de divisão entre $w = 0{,}0064$ e $z = 7312$ resultando num número com $2t$ dígitos. Primeiramente, observamos que o sistema é decimal, com $t = 2$, $e_{min} = -5$ e $e_{max} = 5$. Assim, precisamos armazenar os números dados no sistema indicado: $w = 0{,}64 \cdot 10^{-2}$ e $z = 0{,}73 \cdot 10^{4}$ procedemos com a divisão que nos resulta em $w / z = 0{,}8767 \cdot 10^{-6}$. Como o expoente mínimo é $e_{min} = -5$, o resultado da operação corresponde a um valor menor que o computador é capaz de armazenar, conhecido como *underflow*.

O menor número reconhecido por um computador depende do contexto em que estamos falando. Aqui estão alguns contextos diferentes:

1. **Inteiros (Integers):**

   Para um sistema de 32 bits, o menor número inteiro é geralmente -2,147,483,648.

   Para um sistema de 64 bits, o menor número inteiro é geralmente -9,223,372,036,854,775,808.

2. **Números de Ponto Flutuante (Floating Point Numbers):**

   No padrão IEEE 754 de 32 bits (precisão simples), o menor número positivo diferente de zero é aproximadamente 1.4×10^{-45}.

   No padrão IEEE 754 de 64 bits (precisão dupla), o menor número positivo diferente de zero é aproximadamente 4.9×10^{-324}.

3. **Números em ponto fixo (Fixed Point Numbers):**

   Isso depende da implementação específica do sistema, mas eles são limitados pelo número de bits alocados para a parte inteira e a parte fracionária.

A aproximação de um número $\boldsymbol{y}$ é identificada por $\tilde{y}$

Erro Absoluto $(EA)_y$

Definido pela diferença entre o valor exato e a aproximação: $EA_y = |y - \tilde{y}|$. Mas podemos não saber qual é o valor exato, então, adotamos uma "cota", tal que $\sigma \approx 0$, e assim: $EA_y < \sigma \Leftrightarrow |y - \tilde{y}| < \sigma \Leftrightarrow \tilde{y} - \sigma < y < \tilde{y} + \sigma$.

A aproximação de um número y é identificada por $\tilde{y}$

Erro Relativo $(ER)_y$

Suponha a situação: $w = 50$; $\tilde{W} = 50{,}02$ e $z = 0{,}0004$; $\tilde{Z} = 0{,}0002$. Teríamos que os erros absolutos de cada variável seriam: $EA_w = 0{,}2$ e $EA_z = 0{,}0002$, assim $EA_z < EA_w$ e concluiríamos que a aproximação de z é melhor frente a de w. Mas observe que as grandezas dessas variáveis são muito diferentes, o que nos leva a definir o erro relativo: $ER_y = \left|\frac{y - \tilde{y}}{\tilde{y}}\right|$.

Para o caso citado, teríamos: $ER_y = \left|\frac{50 - 50{,}2}{50{,}2}\right| = 0{,}003984$ e $ER_z = \left|\frac{0{,}0004 - 0{,}0002}{0{,}0002}\right| = 1$, o que nos leva a crer que a aproximação de w é superior à de z.

Erros de arredondamento

Número	*Arrendondamento*	*Truncamento*
2,32	$0{,}232 \cdot 10$	$0{,}232 \cdot 10$
11,054	$0{,}111 \cdot 10^2$	$0{,}110 \cdot 10^2$
-138,17	$-0{,}138 \cdot 10^3$	$-0{,}138 \cdot 10^3$

Propagação de erros

Podemos ter operações matemáticas com mais de uma variável. Assim, temos, além dos erros de representação delas, a propagação de erros causada pelas relações entre variáveis. Para o caso de propagação dos erros absolutos, citam-se os seguintes casos:

a) Soma e subtração entre as variáveis: $EA_{w \pm z} = | EA_w \pm EA_z |$

b) Multiplicação entre as variáveis: $EA_{w \cdot z} = | \tilde{w} EA_z \pm \tilde{z} EA_w |$

c) Divisão entre as variáveis: $EA_{w/z} = \left| \frac{EA_w}{\tilde{z}} - \frac{\tilde{w} EA_z}{\tilde{z}^2} \right|$

Para a propagação dos erros relativos, temos:

a) Soma e subtração: $ER_{w \pm z} = \dfrac{\tilde{w}}{\tilde{w} \pm \tilde{z}} ER_w \pm \dfrac{\tilde{z}}{\tilde{w} \pm \tilde{z}} ER_z$

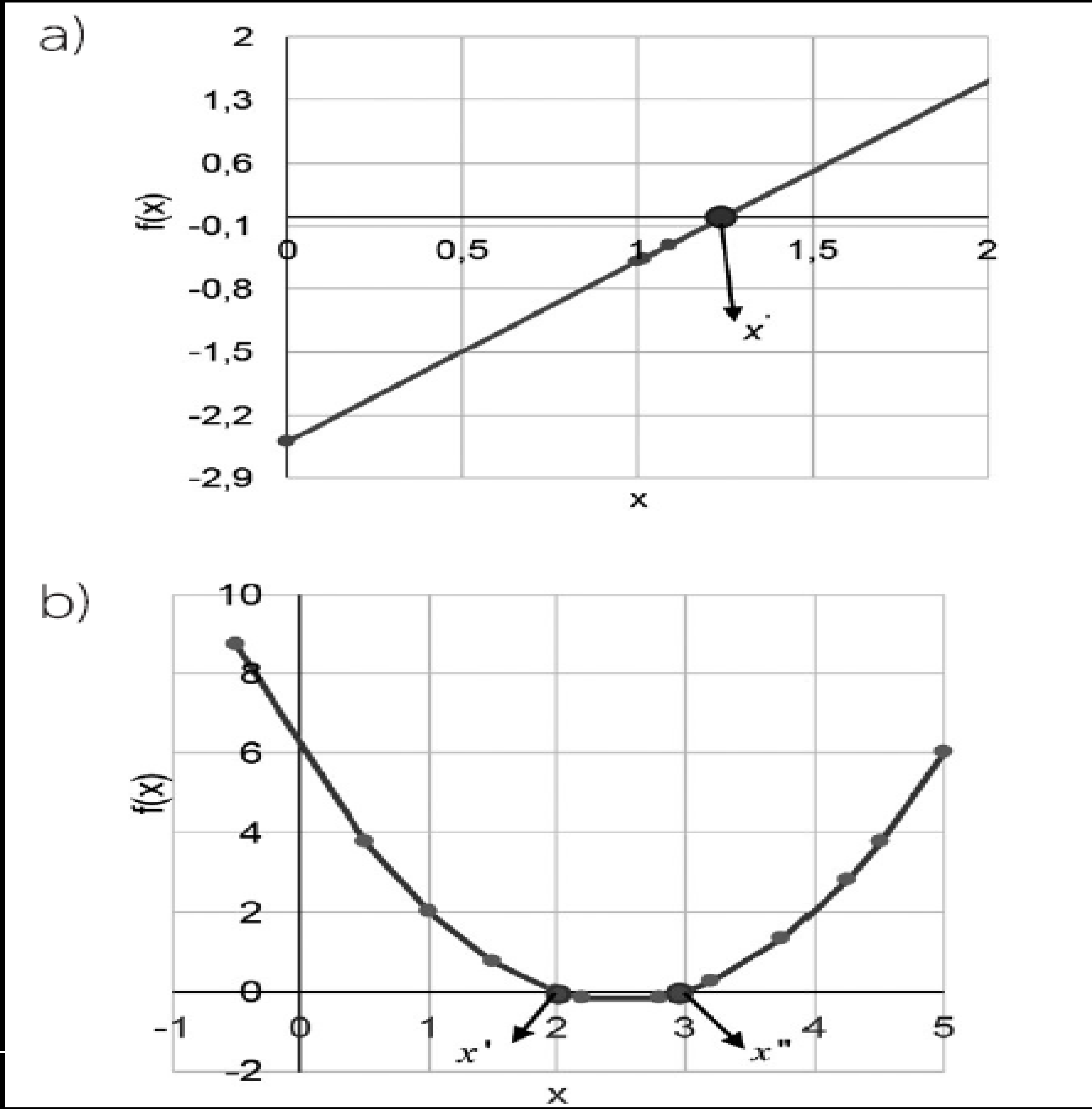
b) Multiplicação: $ER_{w \cdot z} = ER_z + ER_w$

c) Divisão: $ER_{w/z} = ER_w - ER_z$

IMPLEMENTAR FUNÇÃO DE PROPGAÇÃO DE ERROS

75 PTS

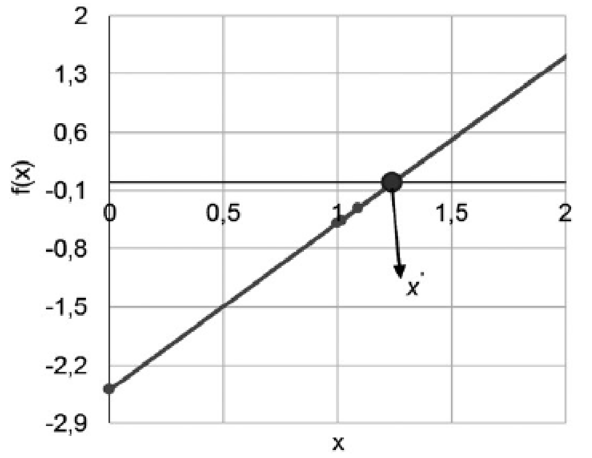
Equações



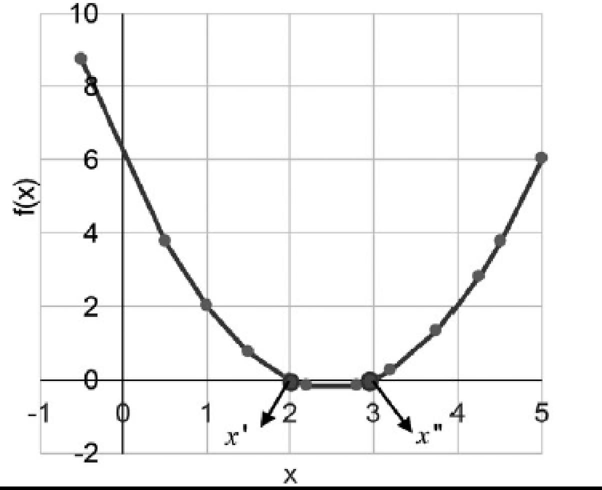
Prof. Dr. Rodrigo Xavier de Almeida Leão

Equações

a)



b)

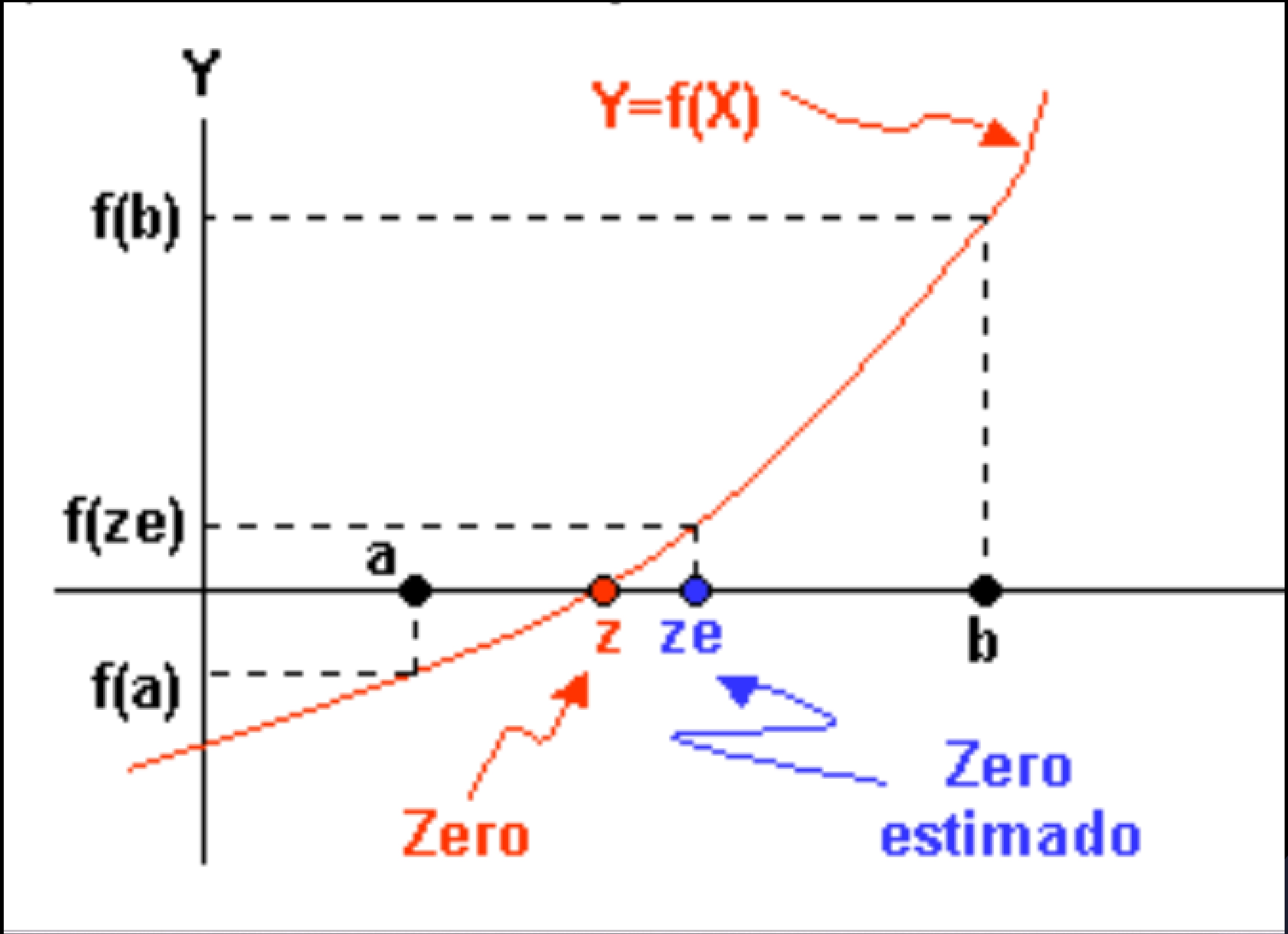


Caso uma função $f(x)$, contínua no intervalo $[a,b]$, possua valores de sinais contrários nos pontos extremos desse intervalo, $f(a) \cdot f(b) < 0$, o intervalo terá, no mínimo, um zero da equação $f(x) = 0$, ou seja, existirá, no mínimo, um número $x^* \in (a,b)$ tal que $f(x^*) = 0$. Verificada a existência de raiz nesse intervalo, precisamos calculá-la por métodos numéricos, que deverão fornecer uma sequência x_k de aproximações, sendo adotado um critério de parada, de modo a cessar o processo iterativo quando for atingido um número predeterminado de iterações, ou se x_k estiver suficientemente próximo do zero da função, ou seja:

$$|f(x_k)| \leq \varepsilon \quad \text{ou} \quad \left| \frac{x_k - x_{k-1}}{x_{k-1}} \right| \leq \varepsilon$$

sendo ε a tolerância estipulada.

Equações



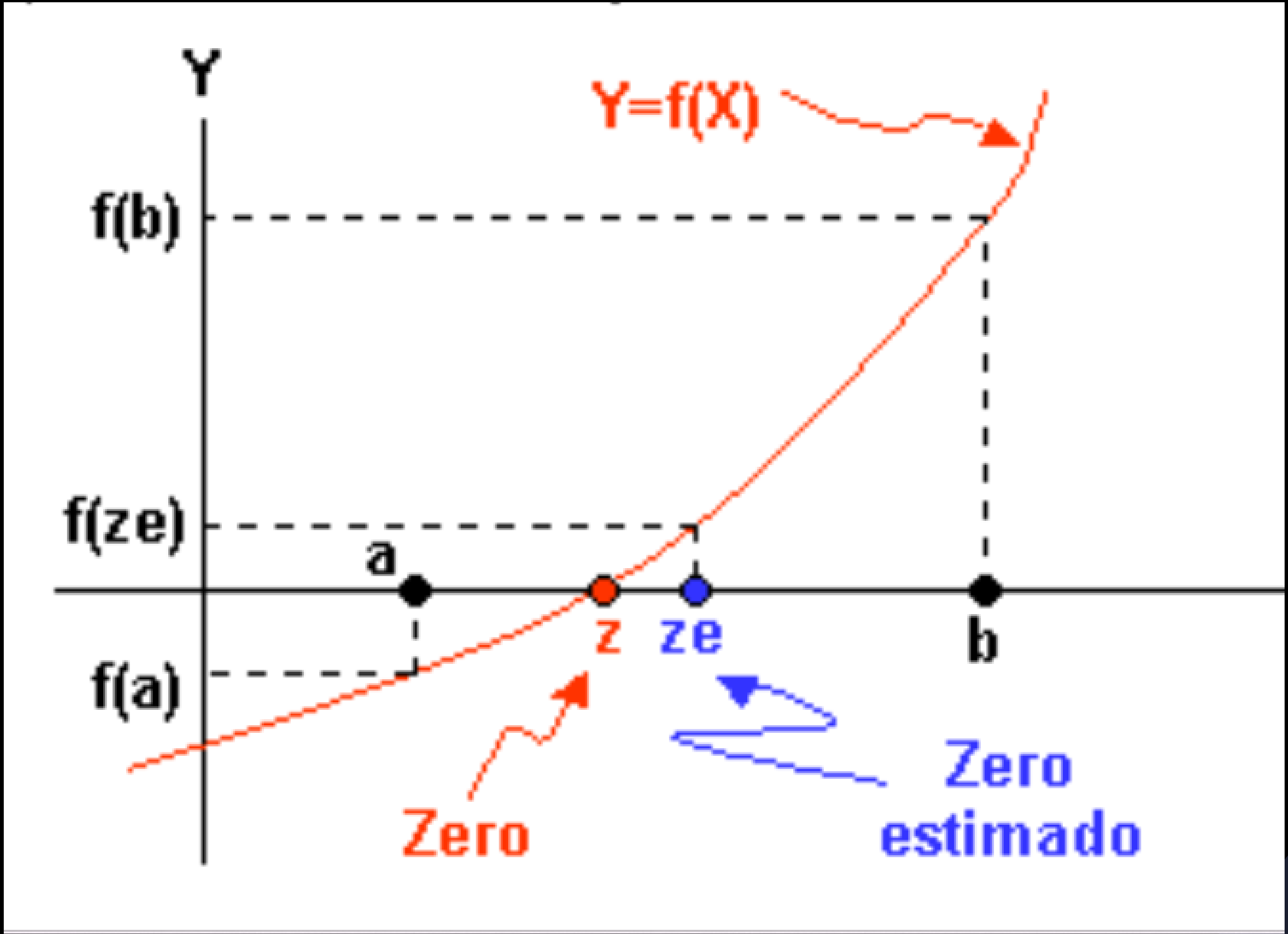
Método da Bissecção

Fundamenta-se na ideia de refinamento de um intervalo inicial, (a,b), que contenha a raiz, de forma iterativa, ao meio. Assim, a cada novo intervalo é atualizado o valor de a ou b de acordo com a função de iteração:

$$x_k = \frac{a+b}{2},\ k = 1,\ 2,\ \ldots$$

Se $f(a)\cdot f(x_k) < 0$, então teremos $b = x_k$, senão $a = x_k$. A desigualdade $f(a)\cdot f(x_k) < 0$ é usada para certificar que haverá pelo menos uma raiz no intervalo (a,b). Assim, a cada nova iteração estamos diminuindo pela metade a distância entre os extremos do intervalo até alcançar o zero de acordo com a precisão desejada.

Equações



IMPLEMENTAR FUNÇÃO PARA SOLUCIONAR EQUAÇÃO DE 2 GRAU COM FÓRMULA DE BHASKARA

50 PTS

IMPLEMENTAR FUNÇÃO DE BUSCA DE ZERO DE FUNÇÕES

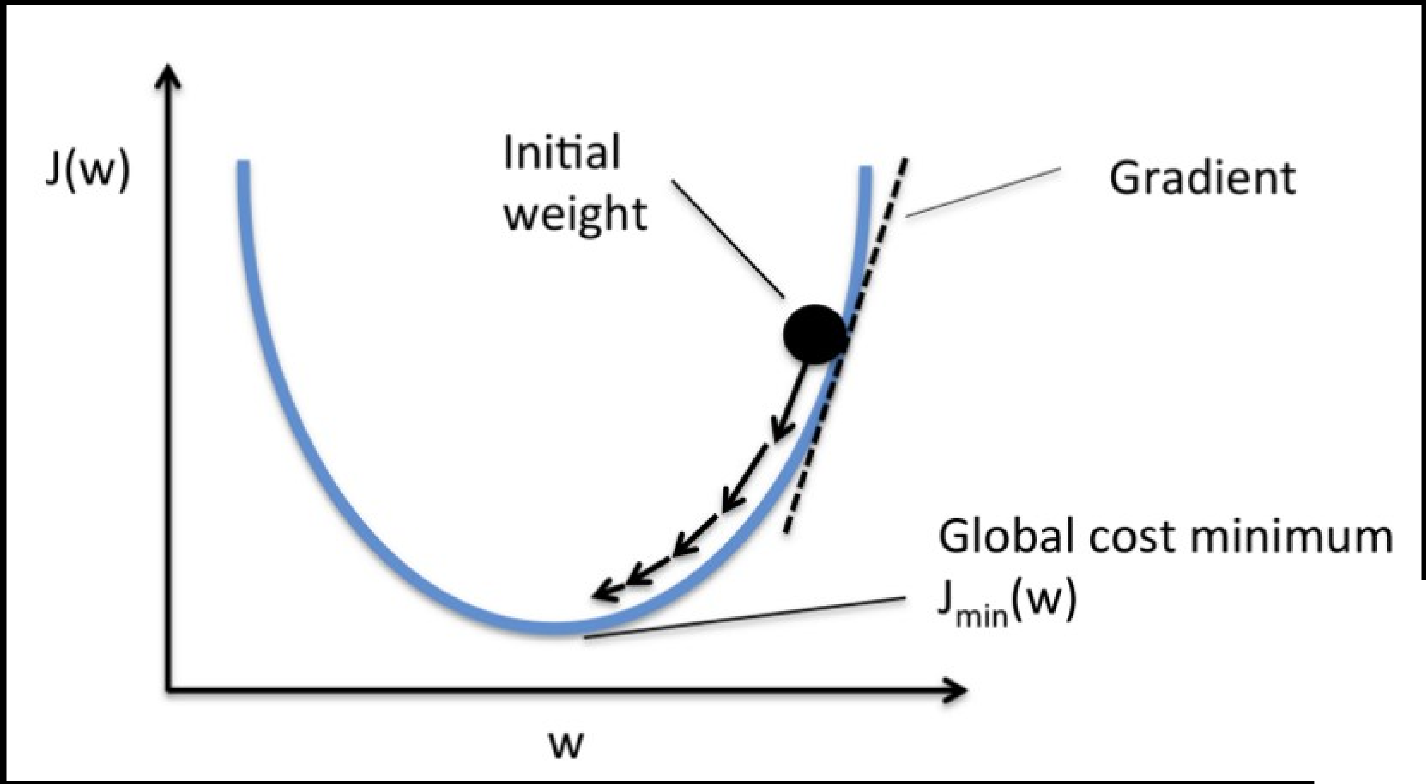
150 PTS

ENCONTRE AS RAÍZES DA EQUAÇÃO

150 PTS

$$x^3 - 2x^2 - 5x + 6 = 0$$

Prof. Dr. Rodrigo Xavier de Almeida Leão


J(w)
Initial
weight
Gradient
Global cost minimum
$J_{min}(w)$
w